RETROSPECTIVE: Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Networks

Hardik Sharma[1] Jongse Park[2] Naveen Suda[3] Liangzhen Lai[3]
Benson Chau[4] Joon Kyung Kim[5] Vikas Chandra[3] Hadi Esmaeilzadeh[6]

[1]Google [2]Korea Advanced Institute of Science & Technology (KAIST) [3]Meta
[4]Georgia Institute of Technology [5]Apple [6]University of California San Diego (UCSD)

hardiksharma@google.com jongse@kaist.ac.kr {naveensuda, liangzhen}@meta.com
ben.chau@gatech.edu jkim62@apple.com vchandra@meta.com hadi@ucsd.edu

I. The Past: Heterogeneous Quantization to Retain Accuracy

In the summer of 2017, when we began this work, the community was amidst the excitement of Deep Neural Networks (DNNs). During that time, both the Machine Learning (ML) and Computer Architecture community had begun grappling with the compute demands and complexity of DNNs. At the same time, the Machine Learning community was in the process of improving DNN accuracy, and each percentage point of improvement mattered. The opportunity to adopt this emerging DNN technology, in the edge device and mobile industry, was alluring. However, the limited power and energy budgets would restrict the complexity of DNN models that could be reasonably deployed on the edge. As a result, simpler and less accurate models were often used instead.

Opportunity for quantization. Enabling the use of state-of-the-art neural networks under stringent constraints had opened up an avenue for innovation. Quantization, the reduction of bit precision for computation and storage, became a focal direction for research. At first, homogeneous quantization was explored, but the initial effort showed a decline in accuracy. Then, the most recent research at the time had shown that DNNs could operate with reduced precision at full accuracy. However, the challenge was that the level of quantization was not homogeneous even within a single neural network and varied across different neural networks. Even now, there is no level of precision that has emerged as the de facto standard for quantization, and the innovations continue (e.g., block floating point [1], Google's bfloat [4], and Microsoft's MX [3]). Thus, a fixed-bandwidth accelerator would either offer limited benefits to accommodate the worst-case bitwidth requirements, or inevitably lead to a degradation in final accuracy. There was a need for flexibility in the architecture to adapt to ML trends that are volatile. Specialized accelerators require significant Non-Recurring Engineering (NRE) cost, making them future proof is essential to accommodate the algorithmic changes.

II. Key Innovation: Composable Architectures at the Finest Granularity Possible, Bits

Making architecture flexibly composable at the bit granularity. We came up with the key idea of this paper at a coffee shop, when we discussed how we could preserve accuracy while enabling ultra power efficiency. The vision was to build bit-level bricks that could dynamically compose and decompose during run-time. We were inspired by two lines of research. The first was Composable Lightweight Processors [O64][1] and Core Fusion [O63] which introduced composability in multicores at the core granularity. The second was Stripes [O2] which explored bit-serial computation.

Our paper[2] introduced and explored the dimension of bit-level flexibility in DNN accelerator architectures, a concept we termed Bit Fusion. This concept dynamically matches bit-level composable processing engines to varying bitwidths that are required by DNN layers. To explore this idea, we designed and implemented a DNN accelerator using a novel bit-flexible computation unit, which we dubbed BitBricks. A 2D array of BitBricks constructs a fusible processing engine that performs the DNN computation at various bitwidths. The encoding and a memory access logic in Bit Fusion stores and retrieves values in the lowest required bitwidth, allowing operands with different bitwidths to be consumed by BitBricks. This strategy proportionally reduces the energy for off-chip memory accesses and increases the effective on-chip storage capacity. Finally, we proposed a block-structured instruction set architecture, called Fusion-ISA that expresses operations of DNN layers as bit-flexible instruction blocks with iterative semantics to amortize the cost of programmability and offer bit-level flexibility at a per DNN layer basis.

By optimizing the memory accesses, on-chip storage, and compute engines at the bit-level granularity, Bit Fusion almost

[1][O..] will refer to the bibliography in the original paper

[2]The artifacts to reproduce all the experimental data and models is open-sourced at https://github.com/hsharma35/bitfusion.

matches the performance of a 250-Watt Titan Xp while consuming less than a watt of power when scaled to the GPU technology node of 16nm.

III. The Present: Bit-Level Optimizations and Composability

Fueling the ML-Architecture Innovation Cycle for Heterogeneous Quantization. Bit Fusion was the pioneer work that demonstrated the *feasibility* of designing an ML accelerator with composability at the finest granularity possible – bits. This architectural feature enabled executing deeply quantized computations with heterogeneous precision needs, a feat necessary to avoid accuracy loss. The work motivated a cycle of innovations – with the architecture community building precision-programmable substrates fueling ML advances, and ML community providing the algorithmic foundations to maximally benefit from such architectural innovations. One the one hand, the paper brought in an ML research direction to architecture community showing that supporting heterogeneous quantization can retain accuracy. On the other hand, it showed the ML community that there is a tangible path to leverage and benefit from innovations in heterogeneous quantization of neural networks below eight bits. There is a swathe of research in the ML community that aims to discover these heterogeneous quantization levels with gradient and non-gradient based approaches.

Analog Neural Processing Units and Bit-Level Composability. Analog and mixed-signal circuits have been a candidate for implementing neural networks from the early days. Deep learning using alternative computing technologies that employ analog-domain logic: (1) using analog multiply-add circuits, or even (2) in-memory computation using Resistive-RAM (ReRAM) or MRAM technologies have thrived in research with multiple prototypes in both industry and academia. Researchers showed that the idea of bit-level composability is useful in coping with the limitations of analog for encoding values and for the high cost of analog/digital conversions. The highly efficient but low-bitwidth analog compute logic could serve as an excellent building block for composable compute units that can fuse and perform higher bitwidth computation while minimizing the cost of crossing the boundary between analog/digital domains. Clearly, analog and mixed-signal designs have potential yet exciting challenges remain to become a prominent alternative to digital CMOS.

Pushing the boundary for composablity in NPUs. Bit Fusion propelled research in bit-flexible architectures that offer various design points optimized for different metrics. The work also provided a novel dimension (bit-composability) in designing NPUs, which researchers combined with other techniques such as bit-serial execution or compression. It also opened the avenue for precision programmability in DNN acceleration by employing it more aggressively at the sub-tensor level or enabling the architecture to dynamically discover and adapt to different quantization levels. Bit Fusion was the first work that explores composability in the context of DNN accelerators and became a prelude for the research that explores the same concept in a different level to support multi-tenancy.

What has stuck with the industry currently? When writing the paper, we focused solely on squeezing out the maximum possible performance-per-watt and area efficiency from the the microarchitectural implementation. However, fully utilizing the compute logic using bit-composable compute units requires wider datapaths compared to the traditional fixed-precision compute units. An alternative approach to precision flexibility has emerged in the industry [2] is to keep the datapath the same width, and simply under-utilize the compute logic for low-precision. Performance for compute in this approach would scale proportionally (so int4 is twice as fast compare to int8, instead of 4x faster in Bit Fusion) when reducing precision. Nevertheless, the benefits in terms of data-movement are attractive, with improved effective bandwidth and on-chip capacity proportional to reduction in precision.

IV. The Future: Virtual Reality and Generative AI

Virtual Reality (VR) and Augmented Reality (AR) workloads introduce a unique challenge – the need for simultaneous acceleration of multiple neural networks that process data from different sensors, execute at different rates, and have varying latency requirements, all within a limited battery budget. Furthermore, the workloads' compute, memory, and even precision demands vary significantly across the co-executing models. As the research and development in the VR/AR domain evolve and the models grow in complexity; composablity and bit-flexibility can be an alluring line of research for best utilizing on-chip resources.

As we also enter the age of Generative AI, the situation seems similar to the time when the paper was written, but in a different scale. There is at present an arms race for advancing capabilities in Natural Language Processing (NLP). One of the challenges being faced is the immense cost of deploying large models due to the huge compute and memory requirements. There is also hesitation with adopting lower-precision NLP models due to the impact of precision on generalizability, accuracy, and user experience is an active research area – all of which affect usability and revenue. Precision-flexible and composable architectures could be of significant importance as they can dynamically adapt to the requirements of Generative AI as the future unfolds.

References

[1] M. Drumond, T. Lin, M. Jaggi, and B. Falsafi, "Training dnns with hybrid block floating point," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[2] NVIDIA, "Nvidia turing gpu architecture." 2018.

[3] B. Rouhani, R. Zhao, V. Elango, R. Shafipour, M. Hall, M. Mesmakhosroshahi, A. More, L. Melnick, M. Golub, G. Varatkar *et al.*, "Shared microexponents: A little shifting goes a long way," *arXiv preprint arXiv:2302.08007*, 2023.

[4] S. Wang and P. Kanwar, "Bfloat16: The secret to high performance on cloud tpus," "https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus", 2019.